Environmental Health Perspectives
Vol. 87, pp. 157–162, 1990

An Index for Cancer Clustering

by Toshiro Tango*

This paper generalizes the index for temporal clustering proposed by Tango in two ways: it allows for nonuniform population distributions across the study period and it is applicable to the detection of disease clustering in space where there are variations in population distribution among categories of the confounding factor such as age and sex. Applications are illustrated with 1833 cases of mortality from uterine cancer in the Tokyo metropolitan area during 1978–1982.

Introduction

The investigation of disease clustering in space, in time, or in both is an important aspect of epidemiological studies in order to find clues to the causative mechanism of the disease in question. For example, the evidence of space-time clustering suggests that individual cases of disease are closely related in both space and time, as is often found in the case with infectious diseases. It has been stated on several occasions that childhood leukemia occurs in clusters in both space and time in many of the studies, which indicates the possibility of viral etiology. Therefore, tests for the detection of space-time clustering have been the subject of considerable research in recent years (*1–6*).

In the study of chronic disease such as cancer, on the other hand, those tests for space-time clustering may not be adequate because cases of chronic disease may be close in space, but they are unlikely to be close in time because of long and variable periods between exposure and diagnosis. Thus, tests for space clustering may be more adequate in this case. However, previous tests for space clustering (*6–8*) have been derived under the unrealistic assumption that the population at risk is fairly uniform across the region. Therefore direct use of those tests would produce spurious evidence of space clustering.

This paper presents a test statistic for the detection of disease clustering in space or in time as an extension of the index C for temporal clustering proposed by Tango (*9*) which can adjust differences in population distribution among categories of the confounding factor such as age and sex. Recently, Whittemore et al. (*10*) proposed a test having the capability of adjusting variations in population distribution among demographic subgroups at different disease risk. However, their procedure, based on the statistic that is essentially identical in form to the index C, is shown to be less adequate than the method proposed in this paper.

*Division of Theoretical Epidemiology, Department of Epidemiology, The Institute of Public Health, 6-1 Shirokanedai 4-chome, Minato-ku, Tokyo 108, Japan.

An Index for Time Clustering

Tango (*9*) proposed an index C for disease clustering in time

$$C = \underset{\sim}{r}^t A \underset{\sim}{r} \tag{1}$$

where $nr^\tau = (n_1, \ldots, n_m)$, $n = n_1 + \ldots + n_m$, denote a vector of observed frequencies in m successive time intervals, which is assumed to be a random sample from the uniform multinomial distribution. Hence, asymptotically,

$$\sqrt{n}(\underset{\sim}{r} - m^{-1}\underset{\sim}{1}) \sim N(\underset{\sim}{0}, m^{-2}V(m\underset{\sim}{1})) \tag{2}$$

where

$$V(\underset{\sim}{x}) = \Delta(\underset{\sim}{x}) - \underset{\sim}{1}\underset{\sim}{1}^t \tag{3}$$

and $\Delta(x)$ is the $m \times m$ diagonal matrix based on the vector $\tilde{x}$ and 1 is the m-dimensional vector of one. The entries a_{ij} of $m \times m$ symmetric matrix A are arbitrary known measures of closeness between ith and jth interval with property $a_i = 1$ and a_{ij} is a monotonically nonincreasing function of d_{ij}, the time distance between ith and jth interval. This index attains its maximum value of 1 if and only if $n_i = n$ for some i and $n_j = 0$ for $j \neq i$. A natural selection for the form of the distance d_{ij} may be

$$d_{ij} = |i - j|. \tag{4}$$

Although the choice of the form of a_{ij} may be variable depending on the situation, an exponential form

$$a_{ij} = \exp(-d_{ij}) \tag{5}$$

has been considered.

The asymptotic distribution function of the index C under the hypothesis of no clustering in time has been, at first, derived using expansion in a series of central chi-square distribution (*9*):

$$Pr\{C < c\} = \sum_{j=0}^{\infty} \alpha_j Pr\{\chi^2_{m-1+2j} < (c - h)/\beta\} \quad (6)$$

where χ^2_g denote the chi-square variable with g degrees of freedom. We shall omit the details on the parameters α_j, h, and β here. However, this formula was not so easy to use in a simple way for more general cases.

Recently, Tango (11) suggested that a better approximation for the distribution of C may be obtained by standardizing C with

$$T = (C - E(C)) / \sqrt{\mathrm{Var}(C)} \quad (7)$$

and approximating it with one central chi-square distribution, i.e., the p-value for the observed value c of the index C can be approximated by

$$Pr\{C > c\} = 1 - I\left(\frac{\upsilon + T\sqrt{2\upsilon}}{2}, \frac{\upsilon}{2}\right) \quad (8)$$

where

$$I(x, \phi) = \int_0^x \frac{1}{\Gamma(\phi)} e^{-t} t^{\phi-1} dt \quad (9)$$

is the incomplete gamma function and

$$E(C) = m^{-2}\{\underset{\sim}{1}^t A \underset{\sim}{1} + n^{-1} tr[AV(m\underset{\sim}{1})]\}, \quad (10)$$

$$\mathrm{Var}(C) = m^{-4}n^{-1}\{4\ \underset{\sim}{1}^t AV(m\underset{\sim}{1})A\ \underset{\sim}{1} + 2n^{-1}tr[(AV(m\underset{\sim}{1}))^2]\} \quad (11)$$

and υ is the degrees of freedom of approximated chi-square distribution and is given by

$$\upsilon = 8[\sqrt{\beta_1}(C)]^{-2} \quad (12)$$

where $\sqrt{\beta_1}\,(C)$ is the skewness of the index C and given by

$$\sqrt{\beta_1}(C) = \frac{8\{3\ \underset{\sim}{1}^t(AV(m\underset{\sim}{1}))^2 A\ \underset{\sim}{1} + n^{-1}tr[(AV(m\underset{\sim}{1}))^3]\}}{\sqrt{n}\{4\ \underset{\sim}{1}^t AV(m\underset{\sim}{1})A\ \underset{\sim}{1} + 2\ n^{-1}tr[(AV(m\underset{\sim}{1}))^2]\}^{1.5}} \quad (13)$$

For convenience in practical applications, the approximated upper 100α percentiles T_α of standardized clustering index T are given in Table 1 as a function of the skewness value $\sqrt{\beta_1}\,(C)$.

Table 1. Approximated percentiles T_α of the standardized clustering index T.

	Nominal α-level		
Skewness[a]	0.05	0.01	0.001
0.0	1.65	2.33	3.09
0.1	1.67	2.40	3.23
0.2	1.70	2.47	3.38
0.3	1.73	2.54	3.52
0.4	1.75	2.62	3.67
0.5	1.77	2.69	3.81
0.6	1.80	2.76	3.96
0.7	1.82	2.83	4.11
0.8	1.84	2.89	4.24
0.9	1.86	2.95	4.38
1.0	1.88	3.02	4.53
1.1	1.89	3.09	4.68
1.2	1.91	3.15	4.83
1.3	1.92	3.21	4.96
1.4	1.94	3.27	5.10
1.5	1.95	3.34	5.25

[a] The entries of this row coincide with the upper percentiles for normal distribution $N(0.1)$.

Extension of the Index

In this section we shall extend the index C so that it is applicable to disease clustering in time or in space where the overall population at risk is not uniform across the region or where there are differences in population distributions among categories of confounding factors such as age.

Let m indicates the number of points in time or in space called regions. Let n_i and E_i ($i = 1, \ldots, m$) denote the observed number of cases and the expected number of cases in the ith region, respectively. Then, as a proper index which can measure the relative intensity of disease incidence or mortality for the ith region, the so-called O-E ratio can be used:

$$\frac{n_i}{E_i} = \frac{\text{observed number}}{\text{expected number}}$$

One example of this quantity is the well known SMR (standardized mortality ratio), which is frequently used in epidemiological studies. Using the above quantity, an extended index can be introduced:

$$G = \sum_{i=1}^{m}\sum_{j=1}^{m} \frac{n_i}{E_i}\frac{n_j}{E_j} a_{ij} = \underset{\sim}{q}^t A \underset{\sim}{q} \quad (14)$$

where a_{ij} is the same form defined by Eq. (5) and d_{ij} may be the Euclidean distance properly scaled between the ith region and the jth region for the case of space clustering problem, E_i can be computed by combining all the regions (i.e., take the standard population to be the entire population being studied), and

$$\underset{\sim}{q}^t = \left(\frac{n_1}{E_1}, \ldots, \frac{n_m}{E_m}\right) \quad (15)$$

In fact, when $E_i = E_j$ for all i,j, then $E_i = n/m$ and

$$G = m^2 C\,. \quad (16)$$

Therefore, it can be said that the index C is reasonably extended to G which can accommodate the variations in the confounding factor distributions over the region. Furthermore, under the hypothesis of no clustering, we have

$$\lim_{n/m \to \infty} E(G) = \underset{\sim}{1}^t A\ \underset{\sim}{1}\,. \quad (17)$$

First, let us consider the problem of disease clustering in time or in space where only the differences are the population size across the region. Let ξ_i denote the population size in the ith region. Then, the vector

$(n_1, \ldots, n_m)$ can be assumed to be a random sample of size n from a nonuniform multinomial distribution with parameter $\underset{\sim}{p}^t = (p_1, \ldots, p_m)$, where $p_i = \xi_i / \Sigma_k \xi_k > 0$, for $i = 1, \ldots, m$. In this case, we have

$$E_i = n\, p_i\,, \tag{18}$$

and, asymptotically,

$$\sqrt{n}\,(\underset{\sim}{q} - \underset{\sim}{1}) \sim N(\underset{\sim}{0}, V(\underset{\sim}{p}_{\text{inv}})), \tag{19}$$

where

$$\underset{\sim}{p}_{\text{inv}} = (p_1^{-1}, \ldots, p_m^{-1})^t. \tag{20}$$

Second, consider the problem of disease clustering in space where the population size is, of course, different over the region with variations in the distributions of the confounding factor such as age.

Let K denote the number of categories in the confounding factor and let ξ_{ik} and n_{ik} denote the population size and the observed number of cases, respectively, for the ith region and the kth category of the confounding factor. Under the hypothesis that there occurs no clustering and the disease incidence rate changes across the categories of the confounding factor, the vector of the observed frequencies $(n_{1k}, \ldots, n_{mk})$ for the kth category of the confounding factor can be a random sample of size $n^{+k} = n_{1k} + \ldots + n_{mk}$ from a nonuniform multinomial distribution with parameter $\underset{\sim}{p}_k^t = (p_{1k}, \ldots, p_{mk})$ where $p_{ik} = \xi_{ik} / \Sigma_j\, \xi_{jk} > 0$, for $i = 1, \ldots, m$ and $k = 1, \ldots, K$. For this case,

$$E_i = \sum_{k=1}^{K} n_{+k} p_{ik} \tag{21}$$

and, asymptotically,

$$\sqrt{n}(\underset{\sim}{q} - \underset{\sim}{1}) \sim N(\underset{\sim}{0}, W), \tag{22}$$

where $W = (w_{ij})$ is the $m \times m$ matrix with element

$$w_{ij} = \begin{cases} n\{\dfrac{1}{E_i} - \dfrac{1}{E_i^2} \displaystyle\sum_{k=1}^{K} n_{+k}\, p_{ik}^2\}, \text{ for } i = j \\[2ex] -\dfrac{n}{E_i E_j} \displaystyle\sum_{k=1}^{K} n_{+k} p_{ik} p_{jk}, \text{ for } i \neq j \end{cases} \tag{23}$$

and $n = n_{+1} + \ldots + n_{+k}$. Needless to say, when $K = 1$, $W = V(\underset{\sim}{p}_{inv})$. Therefore, the mean, variance and skewness values for the index G are shown to be similar in form to Eqs. (10), (11), and (13), respectively, i.e.,

$$E(G) = \underset{\sim}{1}^t A\, \underset{\sim}{1} + n^{-1} tr[AW], \tag{24}$$

$$\text{Var}(G) = n^{-1}\{4\, \underset{\sim}{1}^t AWA\, \underset{\sim}{1} + 2\, n^{-1} tr[(AW)^2]\}, \tag{25}$$

and

$$\sqrt{\beta_1}(G) = \frac{8\{3\, \underset{\sim}{1}^t (AW)^2 A\, \underset{\sim}{1} + n^{-1} tr[(AW)^3]\}}{\sqrt{n}\{4 \underset{\sim}{1}^t AWA\, \underset{\sim}{1} + 2\, n^{-1} tr[(AW)^2]\}^{1.5}}\,. \tag{26}$$

Consequently, the procedure of approximating the asymptotic distribution of the index G under the hypothesis of no clustering can be done exactly in the same way as that for the index C, i.e., we can use the approximation of Eq. (8) where

$$T = \frac{(G - E(G))}{\sqrt{\text{Var}(G)}} \text{ and } \upsilon = 8[\sqrt{\beta_1}(G)]^{-2}. \tag{27}$$

Needless to say, we can use Table 1 to read the approximated upper 100α percentiles of T for the extended index G.

On the other hand, Whittemore et al. (*10*) proposed a test statistic identical in form to the unadjusted index C even for the above-stated situation and approximated it with normal distribution. Clearly, the statistic C itself cannot be a standardized measure. Furthermore, their test has poorer power compared with the test based on the index G since they have used the matrix A as a measure of distance (*11*), and the normal approximation to the asymptotic distribution of the index G should be cautious because it almost always has a substantial amount of positive skewness, which was examined by Tango (*11*) for the detection of time clustering; it will be investigated for the detection of space clustering in detail by simulation study in the next section.

Simulation

To investigate the goodness of approximation by chi-square distribution, we performed the following Monte Carlo simulation. Situations considered here are that there are differences in the overall population size across the region, i.e., $K = 1$.

Step 0: As an entire population Ω, we shall consider the set of 400 points in two dimensional space defined as

$$\Omega = \{X = (u,v) : u = 1, \ldots, 20,\; v = 1, \ldots, 20\}$$

where each point $X = (u,v)$ constitute the centroid of the region.

Then, repeat the following procedure, step 1 to step 3, 100 times.

Step 1: Take random sample with size $m = 100$ (regions) from the set Ω and assume that sampled points $(X_1, \ldots, X_{100})$ constitute the whole region under study. The distance d_{ij} between X_i and X_j is defined as

$$d_{ij} = \sqrt{[(u_i - u_j)^2 + (v_i - v_j)^2]}.$$

Step 2: Take m random numbers from $N(10,2^2)$, say $(r_1, \ldots, r_m)$. Then the value r_i is assigned to ξ_i $(i = 1, \ldots, m)$, the population size for the ith region X_i, and compute

$$p_i = \xi_i \,/ \sum_{j=1}^{100} \xi_j\,.$$

Step 3: For $N = 20(20)100$ and $K = 1$, compute the skewness value $\sqrt{\beta_1}(G)$ and the difference between two kurtosis values, kurtosis value $\beta_2(G)$ and its approximated value $\beta_2(\chi^2_\nu) = 3 + 12/\nu$, where

$$\beta_2(G) = \frac{48\{4\underline{1}^t(AW)^3A\,\underline{1} + n^{-1}tr[(AW)^4]\}}{n\{4\underline{1}^tAWA\,\underline{1} + 2\,n^{-1}tr[(AW)^2]\}^2} \tag{28}$$

The results are given in Table 2 showing that the asymptotic distribution of the index have a substantial amount of positive skewness and that the chi-square approximation is fairly good.

Table 2. Results of Monte Carlo simulation (100 trials) for examining goodness of approximation of the distribution of the index G to a chi-square distribution in the case of space clustering described in the fourth section ($m = 100$ and $K = 1$).

N		Mean	Median	SD	Minimum	Maximum
20	Skewness[a]	0.3989	0.3959	0.0144	0.3707	0.4350
	Difference[b]	0.0426	0.0412	0.0101	0.0262	0.0857
40	Skewness	1.0636	1.0570	0.0712	0.9125	1.3881
	Difference	0.1203	0.1079	0.0852	0.0188	0.6107
60	Skewness	1.0094	0.9981	0.0781	0.8375	1.3585
	Difference	0.0860	0.0507	0.0722	− 0.0327	0.5101
80	Skewness	0.8767	0.8669	0.0686	0.7269	1.1806
	Difference	0.0448	0.0375	0.0544	− 0.0254	0.3840
100	Skewness	0.7841	0.7754	0.0614	0.6502	1.0559
	Difference	0.0358	0.0299	0.0435	− 0.0203	0.3072

[a] Skewness indicates skewness value $\sqrt{\beta_1}(G)$.
[b] Difference indicates the difference between two kurtoses, i.e., $[\beta_2(G) - (3 + 12/\nu)]$, where ν is the adjusted degrees of freedom given in Eq. (27).

Table 4. Results of the application of the clustering index G to 1833 cases of mortality from uterine cancer in 23 Tokyo metropolitan wards during 1978 to 1982.

	Scale parameter λ			
	1	2	3	4
G	28.265	46.907	74.07	103.629
E (G)	26.964	44.217	69.668	97.623
Var (G)	0.655	2.636	7.056	13.451
Skewness	0.318	0.211	0.156	0.126
T	1.607	1.657	1.657	1.638
p-value[a]	≈ 0.05	≈ 0.05	≈ 0.05	≈ 0.05

[a] p-value is read from Table 1.

Application

We shall apply our test to examine the level of clustering among $n = 1833$ cases of mortality from uterine cancer occurred in Tokyo metropolitan area during 1978 to 1982. Table 3 shows that the female distribution of population by age, number of deaths from uterine cancer and its SMR, n_i/E_i, and the latitude and longitude of the geographical centroid in each of $m = 23$ wards. The population numbers for each of $K = 7$ age groups in each of 23 wards were obtained from the 1980 Japanese census. The distance between any two different wards, d_{ij}, was calculated in kilometers using the following approximate formula applicable to Tokyo metropolitan area:

$$d_{ij} = \sqrt{[110.9^2 \times (u_i - u_j)^2 + 90.15^2 \times (v_i - v_j)^2]} \text{ (in km)}$$

where u_i and v_i indicate the latitude and longitude of

Table 3. Female distribution of population by age, number of deaths from uterine cancer, and its standard mortality ratio in each of 23 Tokyo metropolitan wards during 1978 to 1982. Also shown is the latitude and longitude of the geographical centroid of each of wards.

	Age							Deaths from uterine cancer		Geographical centroid	
Ward	20–29	30–39	40–49	50–59	60–69	70–79	80–	Number	SMR	Latitude	Longitude
Chiyoda	3937	4004	4024	4032	3097	1916	610	18	107	35.690	139.755
Chuoh	5749	6637	5865	5569	4802	3072	1021	27	104	35.667	139.775
Minato	17159	19062	15561	13339	9374	5422	1707	49	93	35.654	139.754
Shinjuku	33767	29944	24657	20918	13647	7970	2800	86	108	35.689	139.707
Bunkyo	18123	16253	14219	12745	9251	5590	2019	45	86	35.704	139.756
Taitoh	12853	14595	13808	12493	9813	5940	1839	63	119	35.708	139.784
Sumida	17730	19098	16997	14149	10411	5995	1731	69	122	35.694	139.798
Kohtoh	27272	33613	25922	18536	12006	5752	1521	81	122	35.668	139.819
Shinagawa	31340	29415	25099	21076	14091	8531	2915	63	76	35.606	139.736
Megro	27411	24525	19119	15682	10740	6940	2445	59	91	35.628	139.695
Ohta	54216	56012	46588	38217	25314	14542	4596	153	105	35.576	139.724
Setagaya	76547	67062	55094	43672	27998	17710	6184	161	94	35.642	139.656
Shibuya	26863	21850	17953	15069	9651	5891	2218	62	106	35.660	139.701
Nakano	35776	28712	24083	19490	13072	8180	2896	79	101	35.703	139.666
Suginami	56188	44649	36706	30218	20224	13235	4893	113	91	35.696	139.639
Toshima	28638	24117	20289	17333	11794	7196	2333	77	113	35.729	139.718
Kita	32271	33604	28465	22919	15287	9138	2864	95	107	35.749	139.736
Arakawa	15054	15310	14581	12183	9034	5152	1566	61	124	35.732	139.786
Itabashi	43394	43923	35300	25915	16567	9063	3014	78	80	35.747	139.712
Nerima	46218	47758	43811	30032	17594	9756	3163	85	78	35.732	139.655
Adachi	44201	55543	47823	29540	19716	10558	3035	128	111	35.743	139.803
Katsushika	32210	34701	32481	23147	15511	8432	2483	78	88	35.740	139.850
Edogawa	36605	46306	37039	21917	14650	8046	2334	103	118	35.703	139.871
Total deaths	11	69	238	366	501	436	212	1833			

FIGURE 1. Standardized mortality ratio from uterine cancer in Tokyo's 23 metropolitan wards during 1978–1982.

the geographical centroid of the ith ward, respectively. Maximum and minimum distance was 21.443 km and 1.555 km, respectively. As to the closeness measure, we considered

$$a_{ij} = \exp(-d_{ij}/\lambda)$$

where λ is a scale parameter. Large λ will give a test sensitive to large clusters and small λ will give a test sensitive small to clusters. When $\lambda = 2$, for example, we have

$$G = 46.907,\ E(G) = 44.217,$$
$$\mathrm{Var}(G) = 2.636,\ \sqrt{\beta_1}(G) = 0.211$$

and

$$T = \frac{G - E(G)}{\sqrt{\mathrm{Var}(G)}} = \frac{46.907 - 44.217}{\sqrt{2.636}} = 1.657$$

If we have a good computer program for the incomples gamma function, we can obtain the approximated p-value for the observed value of T from Eq. (8). But here, we shall use Table 1 for simplicity. By referring to the row of skewness = 0.2 (an approximation of 0.211), we can read

$$T_{0.05} = 1.70,\ T_{0.01} = 2.47,\ T_{0.001} = 3.38.$$

Therefore, the p-value of $T = 1.657$ is slightly greater than 0.05, indicating a weak but approximately significant evidence of clustering ($p = 0.05$). Results for several values of λ, summarized in Table 4, are very similar one another. Therefore we can make an inference that some kind of space clustering may have occur for the mortality from uterine cancer during 1978 to 1982 in metropolitan Tokyo. Visual inspection of the map of SMR illustrated in Figure 1 suggests that a clustering occurs in the east of Tokyo such as Arakawa (SMR = 124), Taitoh (SMR = 119), Sumida (SMR = 122), Kohtoh (SMR = 122), and Edogawa (SMR = 118). The result might provide a motivation for further investigation of etiologic clues that may explain the clustering of uterine cancer in this area.

Computing time for these statistics required about 4 min of NEC PC 9801 (VX 21) CPU time using a BASIC computer program that is available from the author upon request.

The author thanks S. Hashimoto for his collaboration in providing data on the population and the latitude and longitude in each of Tokyo metropolitan 23 wards. This work was supported in part by a grant in aid for scientific research (grant no. 62530019) from the Ministry of Education, Science and Culture of Japan.

REFERENCES

1. Pinkel, D., and Nefzger D. Some epidemiological features of childhood leukemia in the Buffalo, N.Y. Area. Cancer 12: 351–358 (1959).
2. Heath, C. W., Jr., and Hasterlik, R. J. Leukemia among children in a suburban community. Am. J. Med. 34: 796–812 (1963).
3. Knox, G. Epidemiology of childhood leukemia in Northumberland and Durham. Br. J. Prev. Soc. Med. 18: 17–24 (1964).
4. Ederer, F., Myers, M. H., and Mantel, N. A statistical problem in space and time: do leukemia cases come in clusters? Biometrics 20: 626–638 (1964).
5. David, F. N., and Barton, D. E. Two space-time interaction tests for epidemicity. Br. J. Prev. Soc. Med. 20: 44–48 (1966).
6. Mantel, N. The detection of disease clustering and a generalized regression approach. Cancer Res. 27: 209–220 (1970).
7. Clark, P. J., and Evans, F. C. Distance to nearest neighbor as a measure of spatial relationships in populations. Ecology 35: 445–453 (1954).
8. Lewis, M. S. Spatial clustering in childhood leukemia. J. Chronic Dis. 33: 703–712 (1980).
9. Tango, T. The detection of disease clustering in time. Biometrics 40: 15–26 (1984).
10. Whittemore, A. S. Friend, N., Brown, B. W., and Holly, E. A. A test to detect clusters of disease. Biometrika 74: 631–635 (1987).
11. Tango, T. Asymptotic distribution of an index for disease clustering. Biometrics, in press.